Phân Tích Dữ Liệu Thực Tế với Python

Bài 9.2: Biểu Đồ Thống Kê với Seaborn

Quang-Khai Tran, Ph.D
CyberLab, 04/2023

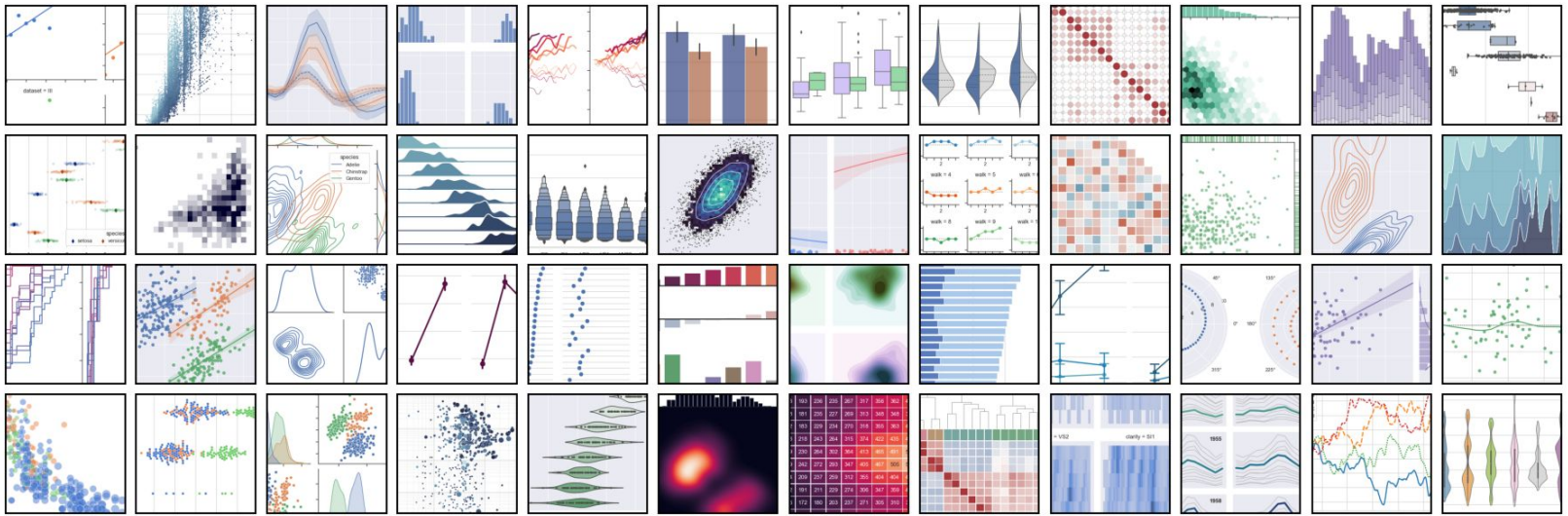
(Ảnh: Internet)

Nội dung

1. **Giới thiệu**
2. **Một số tính năng chính**
3. **Các loại biểu đồ đơn**
4. **Biểu đồ dạng lưới:**
   FacetGrid, PairGrid
5. **Demo: dữ liệu 'nycflights'**
6. **Bài tập & Thảo Luận**

1 Giới thiệu

- Seaborn là thư viện dựa trên matplotlib giúp cho việc trình bày biểu đồ thuận tiện hơn và đẹp hơn, đặc biệt là dữ liệu thống kê



https://seaborn.pydata.org/examples/index.html
https://www.mygreatlearning.com/blog/seaborn-tutorial/

1 Giới thiệu

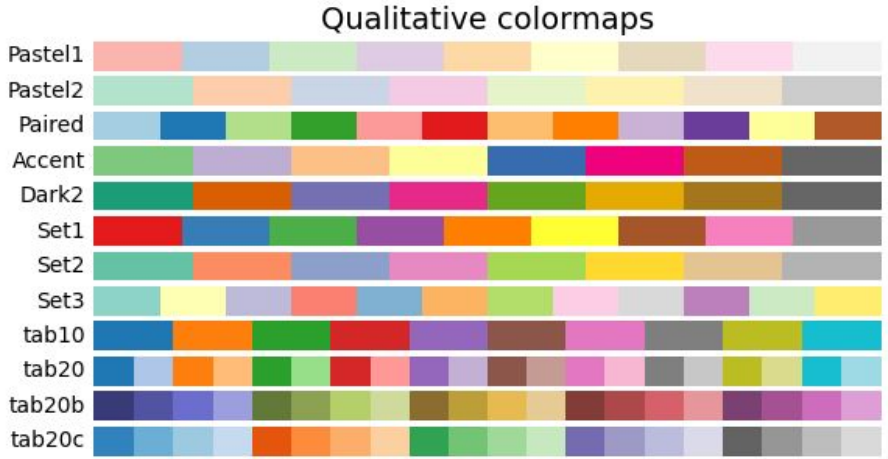
- Một số lợi thế so với Matplotlib
  - Có sẵn nhiều hỗ trợ (theo chủ đề) để việc trình bày biểu đồ được nhanh chóng
  - Đặc biệt là việc trình bày theo ô lưới để phát hiện các mối quan hệ đa biến
  - Hoạt động tốt với DataFrame của Pandas

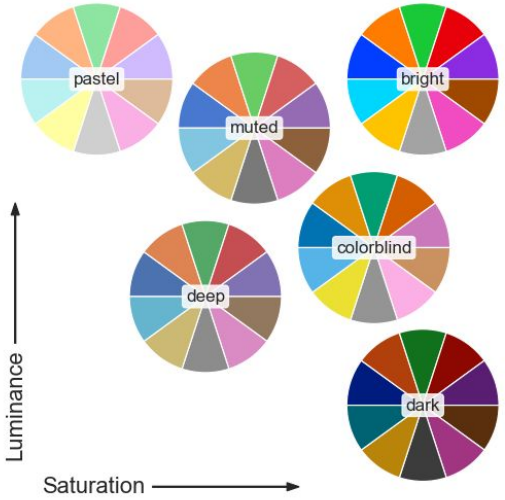
Tham khảo:
https://seaborn.pydata.org/tutorial.html
https://seaborn.pydata.org/examples/index.html
https://www.mygreatlearning.com/blog/seaborn-tutorial/

Tài liệu tiếng Việt:
https://vimentor.com/vi/lesson/1-mo-dau-2
https://ichi.pro/vi/so-tay-tham-khao-cho-30-bieu-do-thong-ke-trong-seaborn-10545983872527

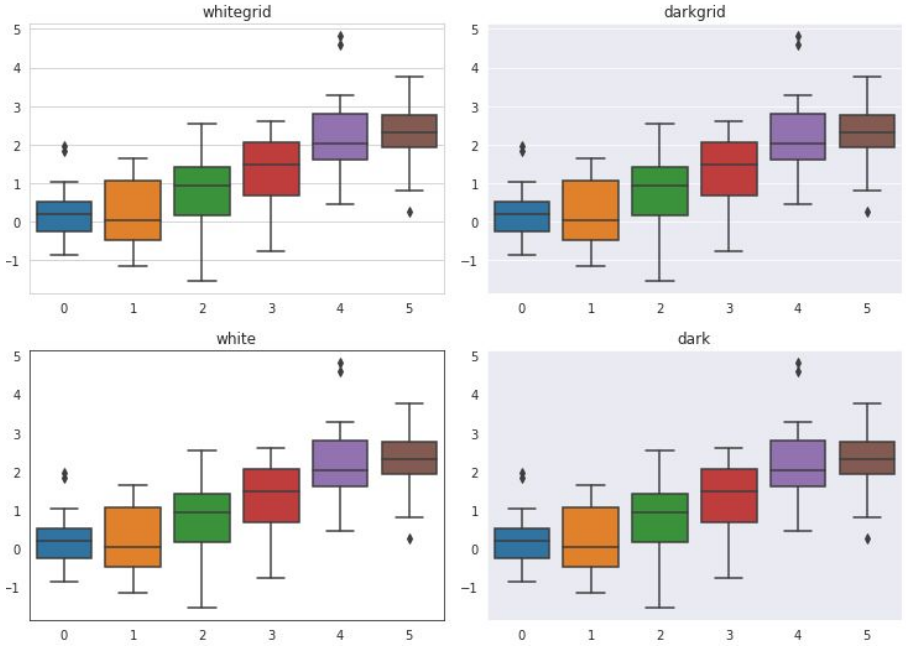
Một số tính năng chính

- **Sử dụng các palette màu sắc**. Tham khảo:
  - https://seaborn.pydata.org/tutorial/color_palettes.html
  - https://matplotlib.org/stable/tutorials/colors/colormaps.html





- **Sử dụng các themes.** Tham khảo:
  - https://seaborn.pydata.org/tutorial/aesthetics.html
  - https://www.codecademy.com/articles/seaborn-design-i

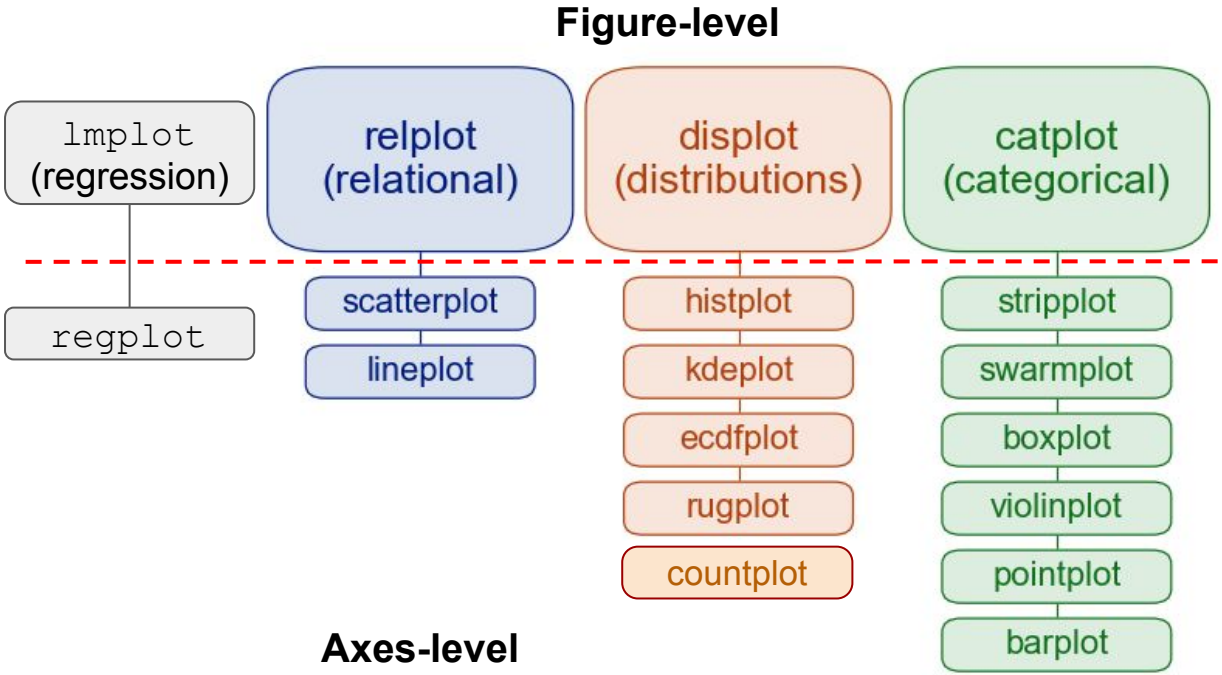


3 Các loại biểu đồ đơn thông dụng

- **Biểu đồ minh họa các loại quan hệ**:
  - Dạng scatter, dạng đường,
  - Dạng linear regression
- **Biểu đồ minh họa phân bố**:
  - Dạng histogram
  - Dạng mật độ
- **Biểu đồ minh họa phân loại**:
  - Dạng cột
  - Dạng box - violin
  - Dạng dải - tổ ong

3 Các loại biểu đồ đơn

❖ Phân cấp các hàm vẽ biểu đồ: figure-level vs. axes-level



Không có: ax = …, color=...

Tinh chỉnh:

- Loại: kind=
- Size: height= / aspect=
- Màu: palette=

3 Các loại biểu đồ đơn

❖ **Vẽ biểu đồ trên các axes**: set tham số 'ax=...' trong câu lệnh vẽ

- Chỉ thực hiện được trên các hàm axes-level

```
# Ví dụ
sns.barplot(data=... , ax=ax1)
sns.scatterplot(data=..., ax=ax2)
```

❖ **Sử dụng tham số 'hue'**:

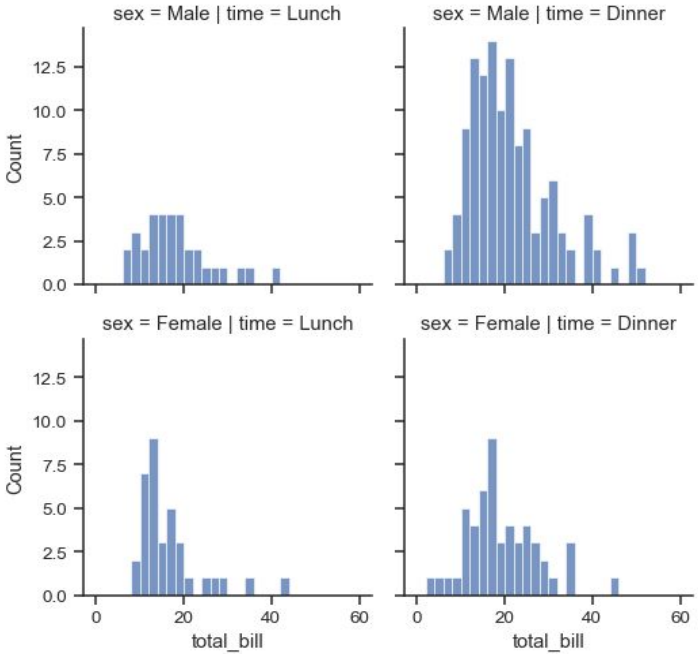
- Phân chia dữ liệu ra các loại khác nhau theo một cột
- Lưu ý: không nên dùng với cột có nhiều loại (5 trở lên)

```
# Ví dụ
sns.barplot(data=... , hue='Giới tính', hue_order=['Nữ', 'Nam'])
```

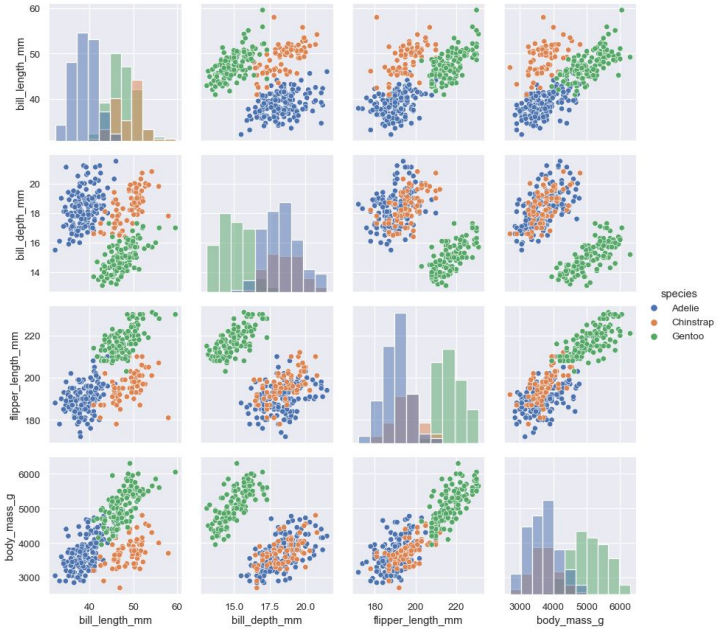
- Kết hợp 'hue_order' để thay đổi thứ tự

4 Biểu đồ dạng lưới

❖ FacetGrid



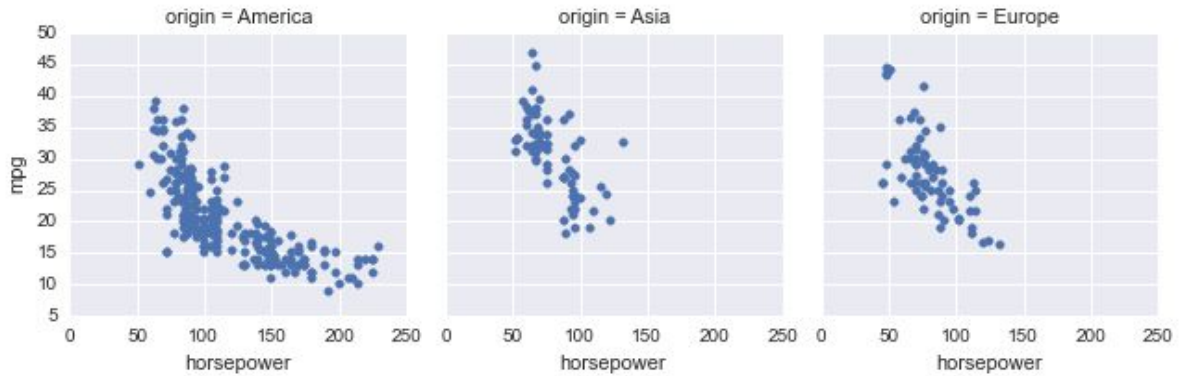
❖ PairGrid



4.1 FacetGrid

- Gồm nhiều biểu đồ biểu thị các mối quan hệ giữa các biến theo phân loại
- Cần xác định 2 loại biến:
  - Các biến phân loại: thường có vài loại (categorical type) nào đó (trong một hay vài cột) để chia dữ liệu thành các subsets
  - Các biến số liệu: mỗi số liệu sẽ được đưa vào các biểu đồ của từng loại

Tham khảo thêm: https://seaborn.pydata.org/generated/seaborn.FacetGrid.html



4.1 FacetGrid

- Cách thực hiện: gồm 2 bước
  1. Khởi tạo một đối tượng biểu đồ dạng FacetGrid
  2. Ánh xạ (mapping) các cột dữ liệu vào biểu đồ:
     - Mapping cơ bản
     - Mapping từ DataFrame

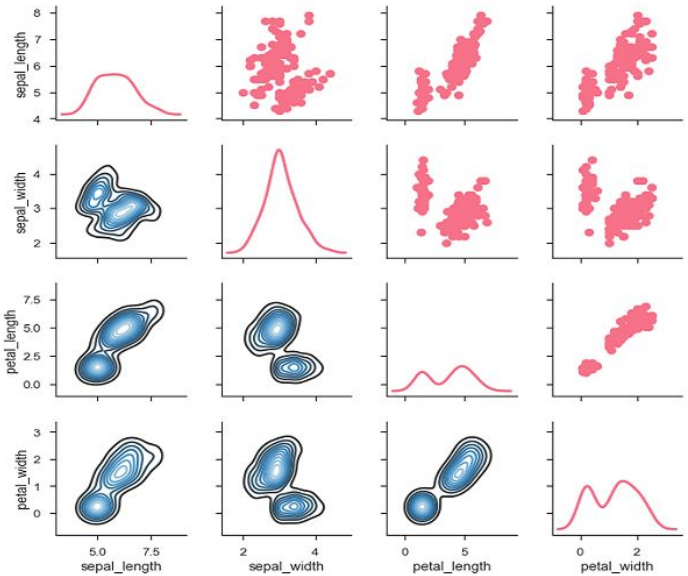
```
bieudo = sns.FacetGrid(df, col='tên1', row='tên2')
bieudo.map(sns.scatterplot, tên3, 'tên4')
bieudo.map_dataframe(sns.histplot, x='tên3', y='tên')
```

Tham khảo thêm:
https://seaborn.pydata.org/generated/seaborn.FacetGrid.html

4.2 PairGrid

- Khác với FacetGrid chia lưới theo phân loại dựa trên cột
- PairGrid chia lưới theo mối quan hệ giữa các cột với nhau

Tham khảo thêm: https://seaborn.pydata.org/generated/seaborn.PairGrid.html



4.2 PairGrid

- Cách thực hiện: có 2 cách
  1. Sử dụng hàm pairplot()
  2. Sử dụng hàm PairGrid và ánh xạ loại biểu đồ mong muốn:
     - Mapping cơ bản
     - Mapping cho các vị trí quanh đường chéo ma trận

```
bieudo = sns.PairGrid(df, corner=True, hue='tên',
vars=[])
bieudo.map(sns.scatterplot)
bieudo.map_offdiag(...)
bieudo.map_diag(...)
bieudo.map_upper(...)
bieudo.map_lower(...)
```

5 Demo: dữ liệu 'nycflights'

Sử dụng Pandas và Seaborn thực hiện phân tích EDA

- Làm giàu dữ liệu bằng cách thêm các thông tin (dẫn xuất) chưa có trong dữ liệu
  - Chuyến bay trễ hay không trễ ⇒ Suy ra từ dep_delay
  - Thông tin về buổi trong ngày ⇒ Suy ra từ thời gian
  - Thông tin về ngày trong tuần ⇒ Suy ra từ ngày tháng năm
- Sử dụng hàm sns.countplot()

6 Bài tập (optional)

Bài tập: sử dụng Pandas và Seaborn thực hiện phân tích EDA cho dữ liệu bán hàng (supermarket_sales_vn.csv)

- Đơn biến (phân tích trên 1 cột):
  - Số đơn: so sánh tổng đơn theo chi nhánh, nhóm hàng, giới tính khách hàng, phương thức thanh toán (cột Payment)
  - Mỗi đơn hàng có một số mặt hàng, vẽ phân bố số lượng đơn theo số mặt hàng (chẳng hạn các đơn hàng có 1,2,3... mặt hàng có số đơn là bao nhiêu?)
  - Thời gian: ngày nào bận rộn nhất tuần? thời điểm nào bận rộn nhất trong ngày?

- Đa biến (cần kết hợp 2 hay nhiều cột):
  - Doanh thu: tỷ lệ tổng doanh thu theo chi nhánh, nhóm hàng, giới tính khách hàng
  - Thuế: tương tự doanh thu
  - Rating: so sánh rating giữa các nhóm hàng, giữa các ngày trong tuần, giới tính
  - Xem xét mối tương quan giữa các biến dạng số

THANK YOU!